f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music

by G. Sayeed Choudhury, Cynthia Requardt, Ichiro Fujinaga, Tim DiLauro, Elizabeth W. Brown, James W. Warner, and Brian Harrington

This paper describes the development of a set of workflow management tools (WMS) that will reduce the manual input necessary to manage the workflow of large-scale digitization projects. The WMS will also support the path from physical object and/or digitized material into a digital library repository by providing effective tools for perusing multimedia elements. The Lester S. Levy Collection of Sheet Music Project at the Milton S. Eisenhower Library at The Johns Hopkins University provides an ideal testbed for the development and evaluation of the WMS. Building upon the previous effort to digitize the entire Collection of over 29,000 pieces of sheet music, optical music recognition (OMR) software will create sound files and full-text lyrics. The combination of image, text and sound files provide a comprehensive multimedia environment. The functionality of the Collection will be enhanced by the incorporation of metadata, the implementation of a disk-based search engine for lyrics, and the development of toolkits for searching sound files.

Contents

Introduction

Digital libraries provide enhanced access and functionality that facilitates scholarly research and education. This realization motivates many libraries to consider large-scale digitization efforts. However, the resources for managing these projects, particularly from the workflow perspective, are significant. This resource issue presents a major challenge for any library considering a large-scale digitization project. The Milton S. Eisenhower Library at The Johns Hopkins University has experienced the impact of this challenge within the context of the Lester S. Levy Collection of Sheet Music Project.

The Lester S. Levy Collection of Sheet Music Project, which represents one of the digital library initiatives of the Eisenhower Library, consists of two phases. With funding from the National Endowment for the Humanities (NEH) for the first phase (Phase One), the Eisenhower Library digitized this collection of over 29,000 pieces of popular American

sheet music that spans the period from 1780 to 1960.

The National Science Foundation (NSF), the Institute for Museum and Library Services (IMLS) and the Levy family has funded the second and current phase (Phase Two). For Phase Two, the Eisenhower Library will create a comprehensive framework of tools that reduce the manual input, and consequently the cost and time, necessary to manage the workflow of large-scale digitization projects. Furthermore, this framework will not only support the path from physical object and/or digitized material into a digital library repository, it will also offer effective tools for perusing the content of the resulting multimedia objects. The Levy Collection of Sheet Music, with its large size and availability in digital format (with significant potential for enrichment), represents an ideal subject for development and evaluation of this framework.

As a result of this project, sound renditions and metadata will be added to the digitized Levy Collection. The sound files and full-text lyrics will be created using optical music recognition (OMR) software, developed by Professor Ichiro Fujinaga of the Peabody Conservatory at The Johns Hopkins University. The search capabilities of the Collection will be enhanced by creating a metadata tool that will allow searching, retrieval, and navigation through the items of the collections and the corresponding text, image, and sound components. These activities will serve the dual purpose of enhancing the Collection's value to scholars and the general public while also creating a large set of multimedia elements that will be used as a testbed for the workflow management tools. To increase the success of the project, the Eisenhower Library has sought appropriate partnerships with faculty and corporations that share common interests.

Impetus for Development

The motivation for the development of this workflow management system (WMS) arises from the lessons learned during Phase One of the Levy Project. The final performance report for Phase One of the Levy Project states "the most useful thing we learned from this project was that you can never overestimate the amount of time it will take to create a quality digital product" (Requardt, 1998). This statement reflects limitations of current workflow software and technology that focus on specific media, rather than functionality, small numbers of objects and use within specific databases. Furthermore, software packages often do not reflect the needs of subject experts.

As libraries convert materials into digital formats, the need for efficient workflow management tools will increase.

As libraries convert materials into digital formats, the need for efficient workflow management tools will increase. The significant human labor required for editing, inspecting, correcting, and "tagging" (with appropriate metadata) digital objects might

inhibit libraries and other organizations from initiating large-scale digitization efforts. By developing the framework of workflow management tools, semi-automated tools will reduce the resource requirements for implementing large-scale digitization projects and provide enhanced search functionality.

Phase One and Goals for Phase Two of Levy Project

During Phase One, the Eisenhower Library created a database of text index records, images of the music and lyrics and color images of the cover sheets from the Levy Collection. This database is available to the general public at http://levysheetmusic.mse.jhu.edu

The Levy Collection consists of over 29,000 pieces of popular American music. While the Collection spans the years 1780 to 1960, its strength lies within its thorough documentation of nineteenth and early twentieth-century America through popular music. The organization of the Collection includes 38 topics such as the circus; dance; drinking, temperance and smoking; fraternal orders; presidents; romantic and sentimental songs; schools and colleges; and transportation.

Besides such famous songs as "The Star-Spangled Banner", "Hail Columbia", and "Yankee Doodle" (all in fine first and subsequent editions), the Collection contains songs and illustrations about many different aspects of American life. Wars, elections, sports, transportation, town and country life, stars of the stage and screen, patriotism, and humor are all described in the verses of the songs and illustrated in the engravings, lithographs, and many forms of early photo reproduction on song covers. An investigation of the Collection reveals much about the history and social life of America.

The expanding lithographic printing industry of the nineteenth century, which coincided with the growth of the sheet music industry, captured on these music covers a vast array of events, fashions, people, and past times, some important and many ephemeral. The Collection provides a social commentary on American life and a distinctive record of their time. Donald Krummel, an authority of American music, described the covers as "probably the best collection of American music chromoliths ever assembled."

The results of Phase One provided users the ability to search the Collection in three modes. First, users can search by subject, a keyword search on the text record. Each of the pieces has been indexed for the subject of the song and/or cover image. Users may also browse the Collection by the topical arrangement of the physical collection. The physical collection's organization scheme includes 38 topics. Finally, users with interest in the graphical elements can examine the Collection by focusing on the cover art.

For Phase Two of the Levy Project, the Collection will be augmented by the addition of sound files (MIDI) along with the full-text lyrics. Both the sound files and lyrics will be created by using optical music recognition (OMR) software that is being developed by Professor Ichiro Fujinaga of the Peabody Conservatory at The Johns Hopkins University. This OMR software will read directly the TIFF images of the sheet music to create the MIDI files and other derivative digital representations of the music (e.g., NIFF, GUIDO). While commercial software exists to create MIDI files, Dr. Fujinaga's offers the following advantages. First, the software runs in batch processing mode, an essential feature given the

large number of sheets of music. Second, the software is platform-independent. Third, the software provides adaptive capabilities and improves accuracy through exemplar-based learning (Fujinaga, 1996). Fourth, the software is open source code. Finally, the software extracts full-text lyrics that will enable an additional method for searching through the Collection.

To further enhance the scholarly value of the sound component of the Levy Collection, a Web-based interface will be developed for a music research toolkit (e.g., David Huron's Humdrum). These toolkits are software tools intended to assist music researchers and can be used for a variety of computer-based musical investigations. This combination of full-text lyrics, images, sound files, and possibly historical video performances, results in a true multimedia environment. Given the diversity of digital content, the results of this project should be applicable across a range of digital library initiatives.

With the addition of multimedia content, it becomes even more important to enhance the users' ability to search, identify, navigate, or browse through a collection of digital objects. The users' must be able to navigate between the image, text, and sound components of individual sheet music items.

With the addition of multimedia content, it becomes even more important to enhance the users' ability to search, identify, navigate, or browse through a collection of digital objects.

To this end, Phase Two of the Levy Project includes multiple improvements to the existing search capability, including a fast disk-based search engine which is important for many parts of the Levy project. For example, with the database of lyrics, we would like to provide full-text searching. Additionally, for an effort to provide "authoritative" names in an unambiguous manner, as explained below, it is important to quickly search the document space for the context of a particular name. These context clues aid in telling one composer from another and provide automated mechanisms for name authority control. In support of these goals, the search engine we are developing will provide both exact string search capabilities, so that users can search for a specific sequence of lyrics, and fuzzy search matches, so that users can search for songs about a general topic.

The retrieval architecture is based on a multi-level token-based inverted index structure, directly indexing the locations of individual annotated word tokens, and allowing rapid access to additional and flexible word classes via secondary hierarchical index structures [1]. The search engine will support unrestricted regular-expression search over sequences of words, word stems, and user-defined word classes, providing efficient query-optimized searches over partially specified lyric patterns. This search capability will be augmented by efficient secondary index-based query expansion to handle word stem, lemma, part-of-speech and thesaurus-based generalizations, as well as efficient search over application

specific word classes definable by the user.

The incorporation of metadata represents another element for improved searching. The extracted full-text of the lyrics will receive a minimal-level text mark-up for presentation, navigation and simple full-text searching. This mark-up will facilitate bi- and multi-directional linking between the index record files containing the first line of lyric or chorus and the areas of corresponding text.

Even though Phase One of the Levy Project focused on digitizing and mounting the sheet music images on the Web, online indexing was created at the sheet music item level. An index record for each piece of music title was created. This record included (when available): the unformatted transcription of title, statement of responsibility, first line of lyric, first line of chorus, dedication, performer, artist/engraver, publication information, plate number, and box and item number. In addition, a controlled vocabulary, in the form of brief subject terms, both for the content of sheet music covers and content of songs, from the Library of Congress' Thesaurus of Graphic Materials. Researchers can search the information which is available as unformatted free text files that can be searched by keyword or phrase.

For the current phase (Phase Two), this "raw material" contained in the index-text files will serve as the basis for creating a powerful metadata tool that will allow searching/retrieval and navigation through the items of the Collection and between the text, image and sound components of the items. The existing index-text will be converted into structured metadata using XML tagging and "bound", similar to a TEI header, with the digital sheet music, along with image, sound and text versions. The XML markup will enable searching between general keyword and precise searches. Name information from the unstructured index-text will be extracted into specific indexes such as composer, lyricist or arranger, and possibly performer, artist, engraver, lithographer, dedicatee, and publisher. Cross-references will direct searchers to index records containing various forms of names, pseudonyms, transformed from the sheet music pieces. All records will contain "authoritative" versions of names and subject terms will also receive mark-up to facilitate subject keyword searching.

In addition to adding structure to the existing data, descriptive metadata will be created to further enhance the records and improve the users' ability to exploit the riches of the Collection, particularly for the sound component. Any enhancements related to sound would be incorporated into the main metadata workflow.

The mark-up of the index-text and incorporation into an overall "sheet music item package" of image, text, and sound will necessarily require some custom effort. To ensure interoperability with other digital library collections, established standards will be adapted whenever possible. Once additional name fields and indexes become established, the Levy metadata will be mapped to the Z39.50 Bib-1 Attribute Set, a process that will contribute toward the desired interoperability. The use of "authoritative" name headings in records and name indexes in Phase Two, as well as the controlled vocabulary from the Thesaurus of Graphic Materials in Phase One, represent examples of widely-adopted standards within libraries. With this interoperability, researchers will be able to exploit Levy metadata to utilize the online Levy Collection in conjunction with other related digital collections or as a pointer to using important related print collections.

Workflow Management System

The enhancements from Phase Two of the Levy Project will result in a digital library of thousands of objects in a variety of media formats that can be used for research, teaching, or even for casual use. Appropriate metadata and toolkits will offer seamless and detailed examination of the digital library. As mentioned previously, the potential of digital libraries is great, but the substantial effort, from a workflow perspective, is equally, if not more, daunting.

The potential of digital libraries is great, but the substantial effort, from a workflow perspective, is equally, if not more, daunting.

In a fundamental sense, the enhanced multimedia collection and search capabilities merely set the stage for testing the workflow management tools that will be developed. By increasing the number and type of digital elements, the Levy Collection becomes an even better testbed.

Figure 1 depicts the workflow management system (WMS) which will be a framework, architecture, and set of software tools that support the path from physical object and/or digitized material into a digital library repository. The WMS will provide the ability to verify and correct links among associated objects, and generate intelligent sampling, abstractions, and derived forms of the data for rapid perusing.

The WMS may be viewed as an accessible cache - a pre-digital library database for preparing and inspecting proposed items to be imported into the digital library. The WMS will include manual, automatic, and semiautomatic aids and processes to support the import process. The accuracy of annotation and the correspondence between image and text will require user participation. However, the generation of derived abstractions or intelligent samples like thumbnails and summarization can be generated automatically or semi-automatically. Mundane functions such as checking that labels remain unique and that each item has a minimum set of metadata and linked data are clearly automatic. While it is unrealistic to expect that WMS will eliminate human intervention, it will reduce the amount of human effort and labor costs, thereby enabling other institutions to undertake digital library projects.

WMS will be an open system providing interfaces and procedures to enable the use of a variety of existing tools within the overall framework and will provide content processing to prepare unstructured data for authoring tools in specific format and dimensions required for application creation. WMS, as an accessible cache, will enable users to select items, choose subsets, tags, process, and present data to application generators. WMS will work with various front-end solutions for viewing, searching, and retrieving objects from the data repository; that is, WMS will be applicable with other digital library projects without constraining choices for digital library repository software (and associated hardware).

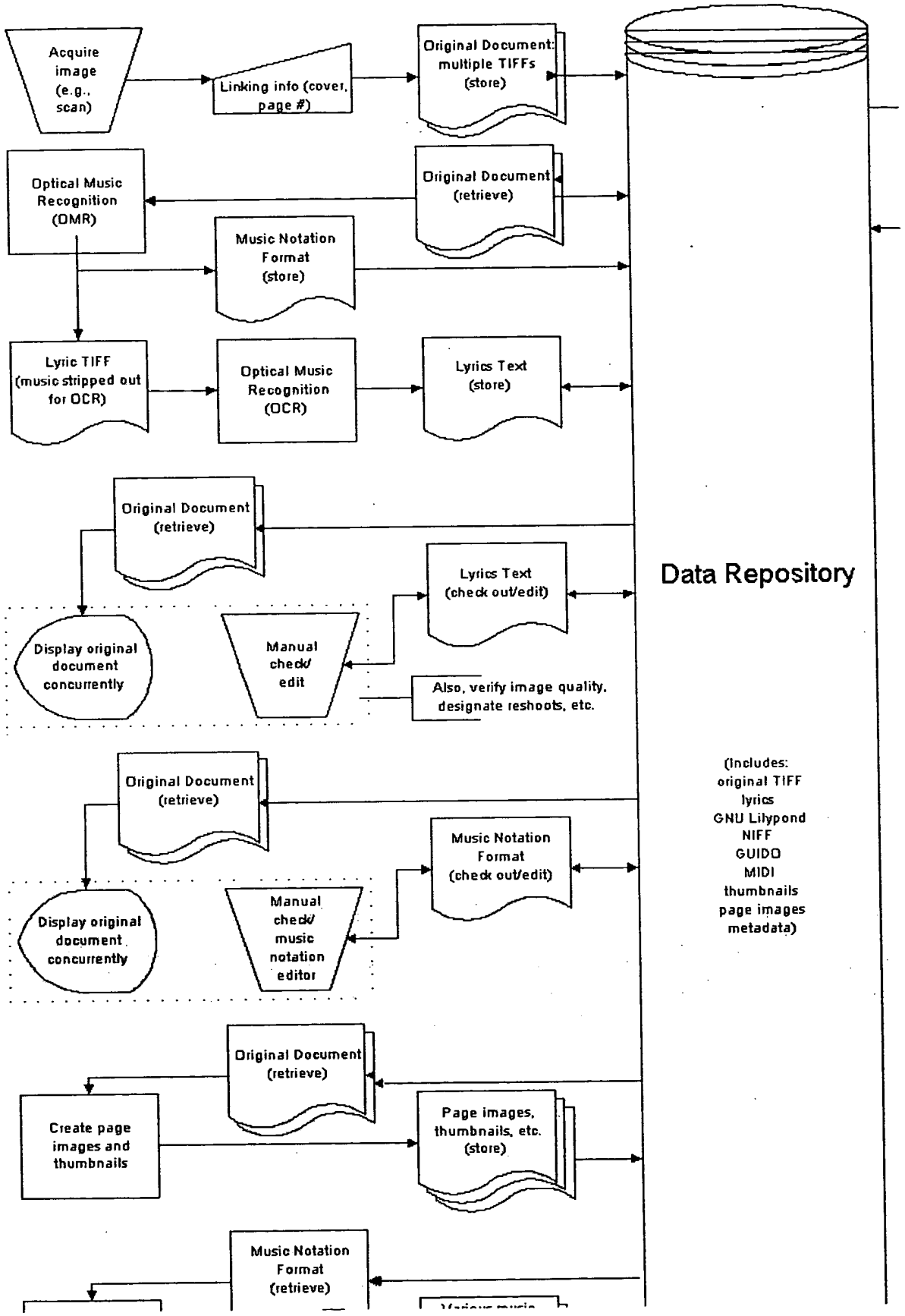
Acquire image (e.g., scan)
Linking info (cover, page #)
Original Document: multiple TIFFs (store)
Optical Music Recognition (OMR)
Original Document (retrieve)
Music Notation Format (store)
Lyric TIFF (music stripped out for OCR)
Optical Music Recognition (OCR)
Lyrics Text (store)
Original Document (retrieve)
Lyrics Text (check out/edit)
Data Repository
Display original document concurrently
Manual check/ edit
Also, verify image quality, designate reshoots, etc.
Original Document (retrieve)
Music Notation Format (check out/edit)
(Includes:
original TIFF
lyrics
GNU Lilypond
NIFF
GUIDO
MIDI
thumbnails
page images
metadata)
Display original document concurrently
Manual check/ music notation editor
Original Document (retrieve)
Create page images and thumbnails
Page images, thumbnails, etc. (store)
Music Notation Format (retrieve)

Figure 1: Schematic of Workflow Management System (WMS)

A comparison of project workflow with and without WMS may prove illustrative. Recall that the proposed OMR software will produce TIFF images of full-text lyrics and either NIFF or GUIDO derivative files. Without the WMS, both sets of files would have to be inspected, edited, marked up, and verified manually using disconnected software tools (which often leads to transferring large files between systems). Ideally, an individual well-versed in music would edit the NIFF or GUIDO files, but these students may not be experts with the specific software used for file processing. This combination of conditions introduces the vast amounts of time and effort for the workflow process. The WMS will provide a comprehensive and integrated set of automatic (e.g., checking for minimum set of administrative metadata), semi-automatic (e.g., creating thumbnails) and manual tools for examining objects in a data repository. That is, individuals can eliminate certain activities (provided through the WMS) and perform various tasks at one workstation. By including connections to other software tools, if an individual feels comfortable using a particular application (e.g., Adobe Photoshop), the WMS can exploit the particular skills of this individual.

Conclusions

The workflow management system (WMS) will reduce the manual input necessary to manage the workflow of large-scale digitization projects. The WMS will also support the path from physical object and/or digitized material into a digital library repository by providing effective tools for perusing multimedia elements. By enhancing the current online Levy Collection database with sound files, full-text lyrics and the ability to search these objects, the online Levy Collection represents an ideal testbed for the development and evaluation of the WMS. The reduction of manual labor for managing the workflow of large-scale digitization efforts will encourage other libraries to move forward with the creation of digital libraries.FM

About the Authors

G. Sayeed Choudhury is the Hodson Director of the Digital Knowledge Center at the Milton S. Eisenhower Library of Johns Hopkins University. Previously, he was the Digital Resources Specialist for the Digital Library Program. Mr. Choudhury serves as principal investigator for sponsored-research projects including the Lester S. Levy Sheet Music Collection Project, funded through NSFs' Digital Libraries Initiative, and the Comprehensive Access to Print Materials (CAPM), funded by the Mellon Foundation.
E-mail: sayeed@jhu.edu

Cynthia Requardt is the William Kurrelmeyer Curator of Special Collections supervising the rare books and manuscripts collections of the Milton S. Eisenhower Library, John Work Garrett Library, and the George Peabody Library. She has directed the projects to digitize the Lester S. Levy Collection, create digital surrogates of medieval manuscripts, and create digital exhibits of rare materials. She holds a masters in history from the University of Maryland and a masters of library science from Syracuse University.

Ichiro Fujinaga has Bachelor's degrees in Music/Percussion and Mathematics from

University of Alberta. He then attended McGill University where he obtained a Master's degree in Music Theory and a Ph.D. in Music Technology.He is currently a faculty member of the Computer Music department at the Peabody Conservatory of Music of Johns Hopkins University.

Tim DiLauro is currently an emerging technologies researcher for the Digital Knowledge Center of The Milton S. Eisenhower Library at Johns Hopkins University. Since 1982, Tim has worked for the University as a Programmer, Systems Programmer, and Sr. Systems Programmer, with a network programming and management component; he has been with the Library for more than 10 of those years. He has also worked as a consultant for several companies with Internet businesses. Over the past five years, Tim's project work has focused on designing systems to improve and simplify user access to information, including the development of access gateways and Web proxies. He is the Library's expert on authentication and access management issues.

Elizabeth W. Brown is Coordinator for the Development of Bibliographic Control at the Milton S. Eisenhower Library at Johns Hopkins University. She coordinates metadata initiatives for the Library's digital library projects, as well as the cataloging of serials and remote-access electronic resources. She manages indexing and metadata production for Project Muse, the joint electronic journals publishing program with the Johns Hopkins University Press. She is also currently involved with the Library's Roman de la Rose medieval manuscript digitization project and Phase II of the Levy Sheet Music Project.

James W. Warner is an undergraduate in the Computer Science department at Johns Hopkins University. He has worked at the Digital Knowledge Center as a student employee for three years. He is a member of the team working on authority file name disambiguation, the lyric database, and the workflow tools for the Levy Project, specializing on information retrieval tasks.

Brian Harrington is the Digital Library Architect for the Milton S. Eisenhower Library of Johns Hopkins University. He has been with the Library for 11 years, working as an Archival Technician, Network Specialist, and Sr. Internet System Administrator. Mr. Harrington has worked on numerous digital library projects, including both phases of the Lester S. Levy Sheet Music Project where he serves as project manager, Project Muse, a joint Library/University Press online journals project and the Roman de la Rose Project, an effort to create a scholarly digital surrogate of medieval manuscripts.

Note

1. For example, the search engine will enable users to distinguish between two documents that contain the word "fly", one in the context of an insect and another in the context of airplanes. Additionally, the search for "fly" would identify all variants of the word. For example, a search for "fly" would identify documents that contain "fly", "flew", "flown", "flies", etc.

References

I. Fujinaga, 1996. "Exemplar-based learning in adaptive optical music recognition system," *Proceedings of the International Computer Music Conference,* pp.55-56.

C. Requardt, 1998. "Preservation of and Automated Access to the Lester S. Levy Collection of Sheet Music," NEH Final Performance Report PS-20945-95 (March).

Editorial history

Paper received 1 May 2000; accepted 10 May 2000.

Contents Index

Digital Workflow Management: The Lester S. Levy Digitized Collection of Sheet Music by G. Sayeed Choudhury, Cynthia Requardt, Ichiro Fujinaga, Tim DiLauro, Elizabeth W. Brown, James W. Warner, and Brian Harrington
First Monday, volume 5, number 6 (June 2000),
URL: http://firstmonday.org/issues/issue5_6/choudhury/index.html